A Semantic-based Method for Visualizing Large Image Collections

Xiao Xie, Xiwen Cai, Junpei Zhou, Nan Cao, Yingcai Wu

Abstract—Interactive visualization of large image collections is important and useful in many applications, such as personal album management and user profiling on images. However, most prior studies focus on using low-level visual features of images, such as texture and color histogram, to create visualizations without considering the more important semantic information embedded in images. This paper proposes a novel visual analytic system to analyze images in a semantic-aware manner. The system mainly comprises two components: a semantic information extractor and a visual layout generator. The semantic information extractor employs an image captioning technique based on convolutional neural network (CNN) to produce descriptive captions for images, which can be transformed into semantic keywords. The layout generator employs a novel co-embedding model to project images and the associated semantic keywords to the same 2D space. Inspired by the galaxy metaphor, we further turn the projected 2D space to a galaxy visualization of images, in which semantic keywords and images are visually encoded as stars and planets. Our system naturally supports multi-scale visualization and navigation, in which users can immediately see a semantic overview of an image collection and drill down for detailed inspection of a certain group of images. Users can iteratively refine the visual layout by integrating their domain knowledge into the co-embedding process. Two task-based evaluations are conducted to demonstrate the effectiveness of our system.

Index Terms—Image visualization, Projection, Interactive refinement

1 Introduction

With the advancement of information technology, images are being created and stored daily on an unprecedented scale. Analyses of large image collections play important roles in a variety of applications, ranging from personal album management, medicine, security, to remote sensing [1]. However, the technologies and tools that empower users to explore and make sense of large image collections are lagging. The recent years have witnessed a growing interest in using visualization methods, such as treemaps [2], node-link diagrams [3], and scatterplots [4], for exploring large image collections. These methods can provide users with a summary of image collections by grouping images based on image similarities, which can be acquired according to intrinsic features (i.e., image pixels and metadata) or user-generated tags. Users are further allowed to drill down to individual images interactively.

Visualization methods have been successfully applied in different systems, such as PhotoMesa [2], PHOTOLAND [5], and ImageHive [6], yet the approaches largely ignore the semantic contents and relationships of objects embedded in the images. The semantic of images can be comprehended as language descriptions of image contents. Semantic information can be crucial in many cases. For instance, the analysis of the semantic content of images posted on social media can reveal the scenes in photos more comprehensively. Such knowledge discloses user preferences, which is valuable in identifying potential targets for advertisements.

Several approaches entail additional information, such as manually produced tags and descriptive text, to analyze the semantic contents of images [7], [8]. However, the information is scarce or even inaccessible in many cases. For instance, personal photo albums and images posted in tweets on Twitter may have few relevant tags and descriptive words. Even if the text descriptions are provided, the images might be inadequately depicted.

The limitation of the existing methods in high-level semantic analysis motivates us to introduce a new method for enabling the semantic-based interactive visualization of large image collections. Nevertheless, semantic-based image visualization is hindered by two major obstacles. The first challenge is extracting the semantic information from images effectively. Low-level information contents, such as objects and their tags, identified from images using image classifications have been exploited to facilitate the exploration and visualization of image collections [7], [8]. Nevertheless, these methods cannot provide sufficient contexts, such as the action and relation of the detected objects, which are important for uncovering insights. The second challenge is visualizing the images with their semantic information. Recent visualizations apply similarity-based methods to project images into 2D space, using visual similarities to organize images. Visual similarities of images are generally refered as the distance between visual features. However, they cannot be directly applied to a semantic based image analysis. Users need to further transform visual appearances into concepts for analyzing semantic contents. Comparatively, semantic similarities, which are based on the distance between language descriptions of image contents, help users cross the gap of conceptualization. Thus, creating multi-scale, intuitive visualization and navigation in a large image collection, such that users can see global and local semantic patterns is significant.

To address the first challenge, we employ an image captioning technique [9] based on convolutional neural network (CNN) [10] that generates reasonable sentence descriptions for images, which can be used to extract semantic information for our method. To address the second challenge, we propose a model of co-embedding images

- *Xiao Xie, Xiwen Cai, Junpei Zhou, and Yingcai Wu are with State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China. E-mail: {xxie,xwcai,zhoujunpei,ycwu}@zju.edu.cn.*
- *Nan Cao is with the College of Design and Innovation, Tongji University, Shanghai, China. E-mail: nan.cao@gmail.com.*

Manuscript received xx; revised xx, 201x.

and the associated semantic keywords, with tailored visual encodings and interactions. We transform the original descriptive captions into semantic keywords to convey different concepts inside the image collection. Galaxy metaphor is employed to create intuitive multi-scale visualizations. Images and associated semantic keywords are represented as different roles of galaxies. This metaphor enables the visualization with smooth transition between the different granularities of images, which endows the visual analytic system with depth and breadth.

The main contributions of this work are as follows:

- A problem characterization and a set of design considerations for semantic-based image visualization.
- A novel model for co-embedding images and words that preserves the complex relationships among images and words.
- A semantic-based image visualization with a tailored multi-scale representation and novel semantic-driven user interactions.

2 Related Works

Our work is closely related to image collection visualizations as well as text visualizations. We discuss the related researches in the following section.

2.1 Semantic Extraction of Images

Image content analysis is the core part of many areas, such as content-based image retrieval. This process aims to disclose underlying semantic content from the pixels of digital images. In this section, we briefly introduce semantic extraction techniques that are most closely related with our work. A complete and in-depth survey is available in [11].

Object classification and detection are the most common techniques for semantic extraction. The rapid development of CV techniques for different neural networks, such as CNN [10] and SOM [12], has recently achieved high accuracies in terms of object detection, thereby enabling users to automatically extract accurate semantic information from images. Semantic information, which is typically represented as keywords, describes concepts embedded in images. However, the results of these methods contain only coarse semantic information. Thus, these approaches encounter the problem of insufficient image semantics. To address this issue, our work uses the neural image caption (NICv2) model [9] to generate sentence descriptions from images. The model evaluation typically shows that the model could generate reasonable sentence descriptions that approximate human-labeled ones on multiple metrics (i.e., BLEU-4, CIDER). Compared with object detection results, sentence descriptions can express objects and their relations, attributes, and activities. Thus, the sentence descriptions of images can be regarded as high-level semantic information.

2.2 Visualization of Image Collection

Image visualizations have been studied over the past years [1]. Previous research has introduced various methods, such as scatterplots [4], treemaps [13] and node-link diagrams [3] to facilitate the image analysis. Liu et al. [14] select representative images and generate a picture collage to summarize the image collection. Crampes et al. [15] analyze social photos that contain personal information and employ a Hasse diagram to show the relationship between photos.

Although the preceding methods facilitate the understanding of images, such methods mainly focuse on utilizing the visual features of images and contained limited semantic information that depends highly on metadata. However, such metadata may be missing or unreliable. Thus, the absence of semantic information substantially hinders users from understanding the images because of the difficulty in interpreting visual features. To integrate semantic information to enhance visualization, Yang et al. [8] use multidimensional scaling (MDS) to project images and apply keywords for annotations and searching. Worring and Koelma [16] visualize images in a pivot table form, thereby supporting the multivariate filtering of images over the user-supplied information and keywords of images.

However, these methods use object detections which can only extract coarse semantic information. To our knowledge, the studies on image visualization along with high-level semantic information are limited. Inspired by previous image visualizations and the importance of introducing serendipity [17] in explorations, we use a projection-based method for two reasons. First, scatterplots are intuitive and easy to read. Second, scatterplots can provide a unified co-embedding space for visualizing words, images, and their similarities and thus clearly show the semantic content embedded in the images. Hence, we develop a novel co-embedding model of images and words to produce a semantic layout of images, thereby showing the latent semantic topics. A multi-scale visual representation based on the galaxy metaphor is introduced with the co-embedding model to enable users to interactively visualize and navigate the galaxy of words and images.

2.3 Visualization of Text

Text visualization has attracted considerable attention [18], [19]. Novel text visualization methods, such as flow-based [20], [21], [22], [23], Wordle-based [24], radial [25], tree-based [26], [27], and projection-based visualizations [28] have been introduced in recent years. ThemeDelta [29] uses line-based visualization to illustrate the convergence and divergence of keywords into different topics. An approach that integrates radial and node-link visualizations is introduced in TopicPanorama [25] to show the full pictures of the relevant topics from multiple sources.

Projection-based methods have also been studied extensively in text visualization [30], [31]. These methods use different techniques, such as the bag-of-words model, to represent documents as high-dimensional vectors that are projected thereafter into a low-dimensional space using various dimension reduction techniques [32], [33], [34]. Several new projection-based methods, such as UTOPIAN [35] and TopicLens [28], have been proposed to leverage the advanced manifold projection technique t-SNE [36] to project documents. In particular, these methods can support interactive visualizations that tightly integrate users into the refinement of the visualization and models.

Although previous projection-based methods have designed reasonable interactive refinement processes, they are mainly designed for documents, without considering the projection of both images and words. Certain co-embedding techniques have been proposed for the projection of heterogeneous data. Canonical correlation analysis (CCA) [37] and correspondence auto encoder [38] have been extensively used in learning common representations of dif-

ferent datasets. Targeting at cross modal retrieval, these techniques emphasize on modeling cross relations and pay less attention to preserving the original similarities. Choo et al. [39] suggest a space alignment method to closely align the related elements from different datasets in the common space. However, the structure of the image data is discriminated from that of the word data, thereby hindering the maintenance of cross-relations while enabling only small deformation of the original space. Therefore, these methods cannot be directly applied in our system. This work develops a novel two-step process of co-embedding images and words. The flexibility of our co-embedding process enables users to refine the projection interactively.

3 Background and System Overview

This section first presents the common tasks of image analyses. Thereafter, we introduce several design rationales derived from the tasks. Lastly, we demonstrate the pipeline and architecture of the system.

3.1 Tasks

Image analysis tasks vary with the application domain. Personal users may want to analyze images to find interesting landscapes for sightseeing. By contrast, social analysts may be more curious about the influence of image contents on tweet propagation. Therefore, we started with a thorough research of image analyses to investigate and extract common tasks in image analysis.

To collect an adequate number of papers, we used an iterative method to review relevant literatures. We sought related works (published in closely related venues such as IEEE VAST, IEEE TVCG, and IEEE Transactions on Multimedia) on IEEE Xplore, ACM Portal, and Google Scholar. In particular, we collected relevant papers through three steps. In Step 1, we searched for relevant papers with several common keywords, such as "image visualization", "image analysis", and "multimedia visualization", and added these papers to a paper list. In Step 2, we selected a paper from the list and looked for other relevant papers from its references. We then removed the selected paper and added newly identified papers to the list. In Step 3, we iteratively repeated Step 2 until no paper was left. In the end, we obtained 32 papers. For task abstraction, we first gathered tasks that were clearly indicated in the papers to form an initial task list. Then, for papers without clear indication, we identified tasks from usage scenarios or case studies and updated the task list. We developed a sufficient task list through this process. We then identified common tasks from the list based on task frequency.

We also conducted brainstorming sessions and held interviews with 14 users (11 undergraduates and 3 graduates, who are all CS majors) to confirm and extend the tasks. Each interview lasted approximately 30 minutes. We summarized four important tasks T1-T4 (T1-T3 from the literature review, and T4 from user feedback) as follows.

T1. **Summarize an Image Collection.** Without clear guidance, finding interesting patterns and contents in images can be difficult [6], [14]. Thus, a simplified image content summary is critical for providing users with access to an image collection. When the content of an image collection is summarized, users can glance at the collection and immediately identify interesting areas.

T2. **Search for Target Images.** Searching for target images is an important task in image analyses [40]. It can be applied to many analysis scenarios. In the medical field, for example, searching for images of one or multiple patients is commonly performed to facilitate patient diagnosis [41].

T3. **Navigate through Images.** Navigation is an alternative method for identifying interesting images [1]. Users typically navigate the image collection when they cannot clearly describe their target images. Navigation is particularly suitable for open-minded exploration that provides high flexibility in user interaction [42].

T4. **Adjust Image Relations.** Image relations significantly impact T1-T3. For example, a system can depend on image relations to generate summary views (T1). Furthermore, such relations enable users to navigate from related images to irrelative images (T3). Searches by examples are also influenced because results are determined based on the relations between the results and the examples (T2). Therefore, the interactive refinement of image relations is necessary.

3.2 Design Rationales

We derive the following design rationales according to the tasks in visually analyzing an image collection.

R1. **Configurable Multi-level Visual Representation.** A hierarchical representation is highly necessary, particularly for handling large-scale image data, for two reasons. First, a multi-level overview of an image collection allows users to immediately see the overall image distribution and quickly identify salient image groups (T1). Second, the visual overview also enables users to drill down to gain further insight (T3). Moreover, the representation should be configurable to account for image relation adjustment (T4) to support the integration of domain knowledge into the visualization layout.

R2. **Image Semantics Revelation.** Revealing image semantics in a system is important and useful in image summarization and exploration (T1-T3). Compared with visual features, such as color and texture, the semantic features of images contain higher abstraction of information. Moreover, these features can be more easily comprehended and accepted by users given that semantics and its similarity (i.e., thesaurus) are naturally embedded into human languages.

R3. **Intuitive Metaphor.** An intuitive visual metaphor can significantly facilitate the navigation, visualization, and understanding of an image collection (T1-T4). In a semantic-based image analysis process, users analyze and explore an image collection through image and semantic relationships. In particular, relationships under consideration include image-image, image-semantic, and semantic-semantic, which are challenging to cope with, particularly for a large image collection. An appropriate metaphor can convey the complex information in an intuitive and easy-to-understand manner.

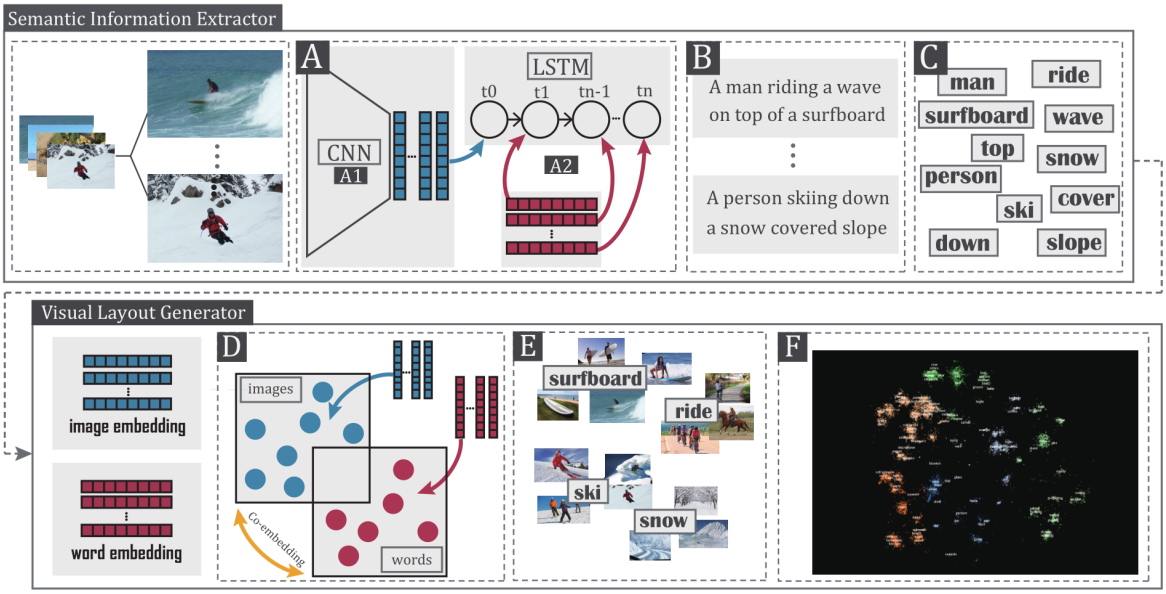


Fig. 1. The system comprises two components: semantic information extractor and visual layout generator. The image captioning model (A), which comprises CNN (A1) and LSTM (A2), translates the raw images to a set of descriptions (B). The descriptions are further transformed into associated keywords (C) for the co-embedding. The co-embedding method (D) processes the image embeddings (in A1), the word embeddings (in A2), and the keywords (in C) to obtain a projection of images and words (E). Based on the co-embedding result, the system produces an interactive visualization of the large image collection (F).

R4. **Flexible Image Query.** The query for images is important for analysts in their evaluation of their hypotheses (T2). However, current methods are rigid because they can only specify the contents of images. Furthermore, these methods fail to provide a compositional query that combines both keywords and sample images. Consequently, a flexible image query mechanism, which includes queries based on keywords, related images, and composition, is required for image analyses.

3.3 System Overview

We develop a web-based system for the exploration of images. For the implementation, we use the Angular-FullStack framework that integrates the MongoDB database, AngularJS frontend, and Node.js server. We employ an image captioning model on the Tensorflow platform and implement our co-embedding model using Node.js C/C++ addon.

Our system comprises two components (Fig. 1), the semantic information extractor and visual layout generator. Our semantic information extractor can translate numerous images into descriptions (Fig. 1(B)) based on the image-captioning model (Fig. 1(A)). We further transform these descriptions into semantic keywords associated with images (Fig. 1(C)). Therefore, we can obtain words that refer to objects and other descriptive words (i.e., running and happy) because the descriptions contain all types of words. To fully utilize the abundant information derived from the semantic information extractor, we develop a novel co-embedding model (Fig. 1(D)) for our visual layout generator. Accepting semantic keywords, image embeddings (blue part in Fig. 1(A1)), and word embeddings (red part in Fig. 1(A2)) as input, the co-embedding model can produce a semantic layout of images (Fig. 1(E)) that can reveal valuable semantic content. Furthermore, we employ a galaxy metaphor to develop an interactive visualization to enhance the exploration of images using the layout (Fig. 1(F)).

4 Model

This section first illustrates the architecture of the image-captioning model and describes how we obtain relevant data for co-embedding images and words. Thereafter, we introduce a two-step method for the co-embedding model. A forest-based method is developed to characterize the semantic relation of words and images in multiple tree structures based on the image captions.

4.1 Image Captioning Model

Our semantic information extractor (Fig. 1) is based on the NIC image captioning model [9]. The model takes raw images as input to generate sentence descriptions from the images. Two state-of-the-art machine learning architectures are involved, namely, CNN for image processing (Fig. 1(A1)) and LSTM for natural language processing (Fig. 1(A2)). CNNs have been widely used in object detection and classification [43], [44]. CNNs can extract robust image features that can be fed into other models through multiple neuron layers. LSTM is a special form of recurrent reural network (RNN). LSTM has a distinct architecture that can address the long-term dependency problem of traditional RNNs. LSTM has shown high capability in accomplishing sequence tasks, such as machine translation and automatic speech recognition.

We use the MSCOCO dataset [45] as the training dataset. The MSCOCO dataset contains 82,783 images of training and 40,504 images of validation, with each image having five human-labeled captions. In short, the image captioning model attempts to learn from the correct captions and translate an image into a human-readable sentence. We denote the parameters of the model with θ, the input image with I and the output sentence with $S = (s_0, \ldots, s_N)$ where s_i represents the ith word in the sentence. The matrix of all

word embeddings are denoted by W_e. The objective of the model can be formulated as:

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta) \quad (1)$$

where p is estimated as follows

$$x_0 = CNN(I) \quad (2)$$

$$x_i = W_e s_i, \quad i \in \{0, \ldots, N-1\} \quad (3)$$

$$p_{i+1} = LSTM(x_i) \quad (4)$$

Consequently, the loss function is the sum of the negative log likelihood of predicting the correct word at each step.

$$L(I,S) = -\sum_{i=1}^{N} \log p_i(S_i) \quad (5)$$

At each step, LSTM accepts an embedding vector of a word ($W_e s_i$) as input (red vector in Fig. 1(A)) and produces a probability distribution of words as the prediction of the next word. Particularly, as illustrated in Fig. 1(A), the image embedding I is inputted into LSTM only at the first step. This step by step inference process enables the model to create multiple sentences for an image with different probabilities. For simplicity, we only selected the sentence with the highest probability as the final caption. Although word embeddings have the same vector size with image embeddings, they are generated by a different model, namely, the skip-gram model [46]. Since word embeddings are important to our co-embedding process and image query, we briefly introduce the skip-gram model.

Training of the Skip-gram Model. The skip-gram model learns a high-quality vector form as the representation of words. Given a word of a sentence, the model considers the nearby words within distance n of the input word as its neighborhood words. While training, the model tries to predict the neighborhood words of the input word. This training procedure ensures that the words with similar context are close in the embedding space. This word representation encodes the semantic patterns of words and is considered a state-of-the-art technique for word embedding.

Characteristic of the Skip-gram Model. The word embeddings from the skip-gram model has an important property: it remains the linear algebraic structure of word meanings. For example, the vector of "man" is the nearest vector of the "King" - "Queen" + "woman" result, and "swimming" is the nearest one of the "walking" - "walked" + "swam" result. This means that word embeddings can maintain certain semantic relationships, such as male-female, verb tense and even country-capital relationships between words. We sum up all the vectors in a caption to represent the corresponding caption because of this property. From this method, we can treat the captions and words in the same way.

4.2 Forest-based Co-embedding of Images and Words

We demonstrate our model for co-embedding images and associated semantic keywords. Given an image set with n pictures $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$, we denote the caption of image I_j as C_j. To obtain the associated semantic keywords, we filter stop words and synonyms from the captions. Then, we collect all the words in a set $\bigcup_{j=1}^{n} C_j$ and denote this word set with m words as $\mathcal{W} = (W_1, W_2, \ldots, W_m)$.

Based on literature reviews [47], in map or galaxy metaphor based visualization, the distance between objects can be used to represent the similarity. Therefore, we determine that the co-embedding result should have two important properties to provide a meaningful semantic layout. First, words should be close to related images, thereby serving as a good annotation. Second, the images that share similar semantic information should be close to each other for the convenient comprehension of semantic content in images. Thus, we employ a two-step co-embedding process. A two-step method is selected over a one-step method because of two reasons. One is the difficulty in considering the multiple relationships between images and images, words and words, and images and words at the same time. In our preliminary studies, we have developed a new cost function for t-SNE to preserve these relationships. However, these relationships, always interfering each other when they are considered together, make it hard to tune the model. Dealing with them separately in two steps makes it easier to generate reasonable layouts. The other is the ease of controlling the modification of the model for users. By processing the multiplex relationships separately, we can provide users with a flexible interface through which the multiplex relationships can be modified without disturbing each other. Hence, users can fully leverage their knowledge on different relationships and refine the layout through interactions.

We expect the final layout to preserve the semantic relation of images. Thus, we deal with the relations between images and images and the relations between images and words, and then embed both images and words into a semantic space. The first step is to obtain the local semantic structures of images, which guarantees that words are placed close to their related images. The second step is to reconstruct images in the semantic space, which cluster the images with similar semantic information. We begin with describing the pre-processing of data and illustrate our co-embedding method sequentially.

Pre-processing. Before the co-embedding process, we use t-SNE to embed images and words respectively as shown in Fig. 2(A). t-SNE is used because it generates more reasonable dimensionality reduction results than the other methods in our experiments. We denote the distance between images as $d(I_j, I_k)$ and the distance between words as $d(W_j, W_k)$. Following previous image processing methods [36], we use $Euclidean$ distance to compute $d(I_j, I_k)$. For $d(W_j, W_k)$, we use $cosine$ distance because word2vec [46] also applies $cosine$ distance as the distance metrics. We denote the 2D embedding of image I_j as P_{I_j} and the 2D embedding of word W_i as $P^*_{W_i}$. Then, we normalize the 2D embeddings of the images and words, respectively. For simplicity, we use $\mathcal{P}$ to represent the 2D embedding space of images and $\mathcal{P}^*$ to represent that of words.

4.2.1 Obtaining Local Semantic Structures

In this step, we produce a preliminary co-embedding by embedding words into $\mathcal{P}$ and obtain the local semantic structures of images. A local semantic structure is a group of images with similar visual features and semantics. First, we construct a bi-directional bindings of images and words. Then we embed words into $\mathcal{P}$. Finally, we extract m trees interpreted as the local semantic structures of images as shown in Fig. 2(D), which are preserved for the next step.

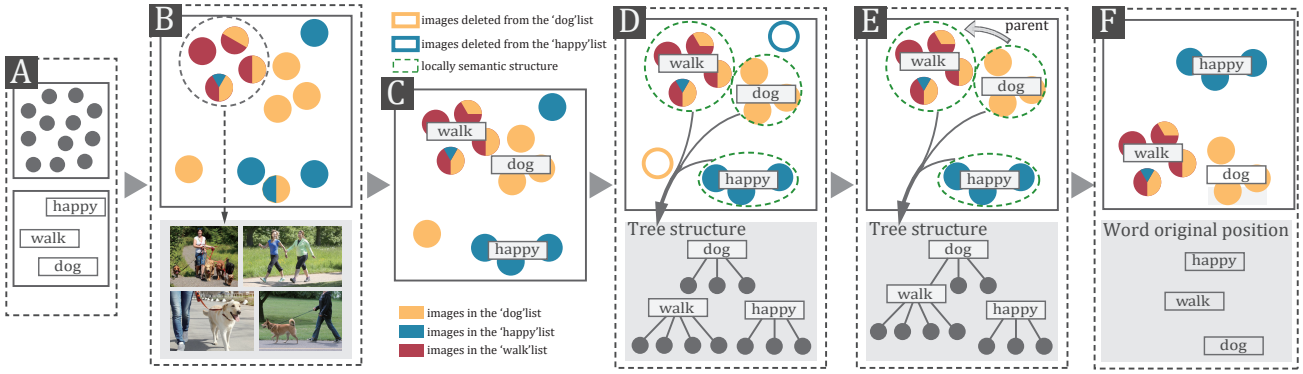


Fig. 2. (A) Project the images and words separately. (B) Assign the images to words according to their similarities. (C) Calculate the centroid of each image collection as the position of the corresponding word. (D) Delete the outlier and recalculate the centroid until all images satisfy the threshold, and then obtain the locally semantic structure. (E) Determine the relations between words according to the confidence. Here the confidence between dog and walk is 0.75 while that of happy and walk is 0.25, which can be calculated by (14). (F) Reconstruct the image according to the relationship between words using the locally semantic structure kept.

Binding words and images. We attempt to bind images and words in a bi-directional manner. Through the binding process, each word can find related images and each image can find related words. This process can help us build the relation between words and images.

As shown in Fig. 2(B), each image represented by a pie is painted by one or more colors. The colors denote which words are related to a certain image. For example, the pie with three colors depict that this image share a high similarity with both "dog" and "walk" and a lower similarity with "happy". Given this compositionality, we define the similarity, $Simi(W_i, I_j)$, between W_i and I_j as:

$$Simi(W_i, I_j) = 1 - \min_{W_k \in C_j} d(W_i, W_k) \tag{6}$$

We define the word-image similarity because we find the caption model occasionally mistakes an object for another object similar in meaning (e.g. dogs and cats are sometimes mistaken for each other). With the similarity, users can adjust the result of captioning through interaction when the result of the caption model is unreliable. Therefore, for a specific word W_i , we define the set of its related images as $\mathcal{I}_{W_i}$ and compute it as:

$$\mathcal{I}_{W_i} = \{I_j \mid I_j \in \mathcal{I},\ Simi(W_i, I_j) \geq MinSimi\} \tag{7}$$

where $MinSimi$ is the threshold for the minimum similarity, which is defaulted to 1.0, its maximum value. When $MinSimi$ is 1.0, $\mathcal{I}_{W_i}$ only contains the image whose caption contains W_j. Similarly, for each image I_j, we define the related words $\mathcal{W}_{I_j}$ as:

$$\mathcal{W}_{I_j} = \{W_i \mid W_i \in \mathcal{W},\ I_j \in \mathcal{I}_{W_i}\} \tag{8}$$

Hence, we use $\mathcal{I}_W$ and $\mathcal{W}_I$ to represent the relation between images and words.

Embedding words. After detecting the relations between images and words, we conduct a preliminary co-embedding that preserve the relations. We expect each word to be embedded in a place close to its related images. Embedding word W_i into $\mathcal{P}$ can be described as minimizing the sum of the weighted distances of W_i to the related images, which can be expressed as:

$$P_{W_i} = \arg\min_{P} \sum_{I_j \in \mathcal{I}_{W_i}} Simi(W_i, I_j)||P_{I_j} - P|| \tag{9}$$

where P is any position in 2D space $\mathcal{P}$. The problem-solving process is similar to finding the geometric median of a set of points (Fig. 2(C)), whose approximate solution can be found with the gradient descent. However, it might lead to the result that some of the images in $\mathcal{I}_{W_i}$ are far from W_i. Thus, we iteratively remove these images from $\mathcal{I}_{W_i}$ and recalculate the position of W_i, to find its optimized position according to a user-defined threshold $MaxDist$, which controls the maximal distance between word W_i and the images in $\mathcal{I}_{W_i}$. For each iteration, we find the image I_f which is furthermost from W_i by:

$$I_f = \arg\max_{I_j \in \mathcal{I}_{W_i}} ||P_{W_i} - P_{I_j}|| \tag{10}$$

If $||P_{W_i} - I_f|| > MaxDist$, we delete I_f from list $\mathcal{I}_{W_i}$ and then recompute the position of W_i according to (9). As shown in Fig. 2(D), the blue pie, which is too far from "happy", and the yellow pie which far from the "dog", are deleted. We repeatedly delete the images in (10) until $||P_{W_i} - P_{I_f}|| \leq MaxDist$ is satisfied. The deletion is a bi-directional process, that is, when image I_f is deleted from $\mathcal{I}_{W_i}$, word W_i is also deleted from $\mathcal{W}_{I_j}$. Through this iteration, we simplify the complex relation between words and images, to find the local semantic structures of images.

Extracting local semantic structures. We have simplified the relation between words and images. However, preserving the relations between an image and multiple related words can break visually similar images into different groups, thereby discarding the information from visual features. To preserve the information provided by visual features, we need to further find the most related word for each image to detect the local semantic structures of the images (Fig. 2(D)). To illustrate, we construct a pair of values (S_i, D_i) for each word W_i in $\mathcal{W}_{I_j}$ where $S_i = Simi(W_i, I_j)$ and $D_i = ||W_i - I_j||$. After sorting, word W_j ranks ahead word W_k must satisfy $S_j < S_k$. If $S_j = S_k$ holds, $D_j < D_k$ should also hold. If $\mathcal{W}_{I_j}$ is empty, I_j has no parent nodes. Otherwise, the first word in $\mathcal{W}_{I_j}$ is set as the parent node of I_j. In this manner, we can obtain the local semantic structure for each word, which contains semantically and visually similar images, and represent the local semantic structures in a forest form for the next step.

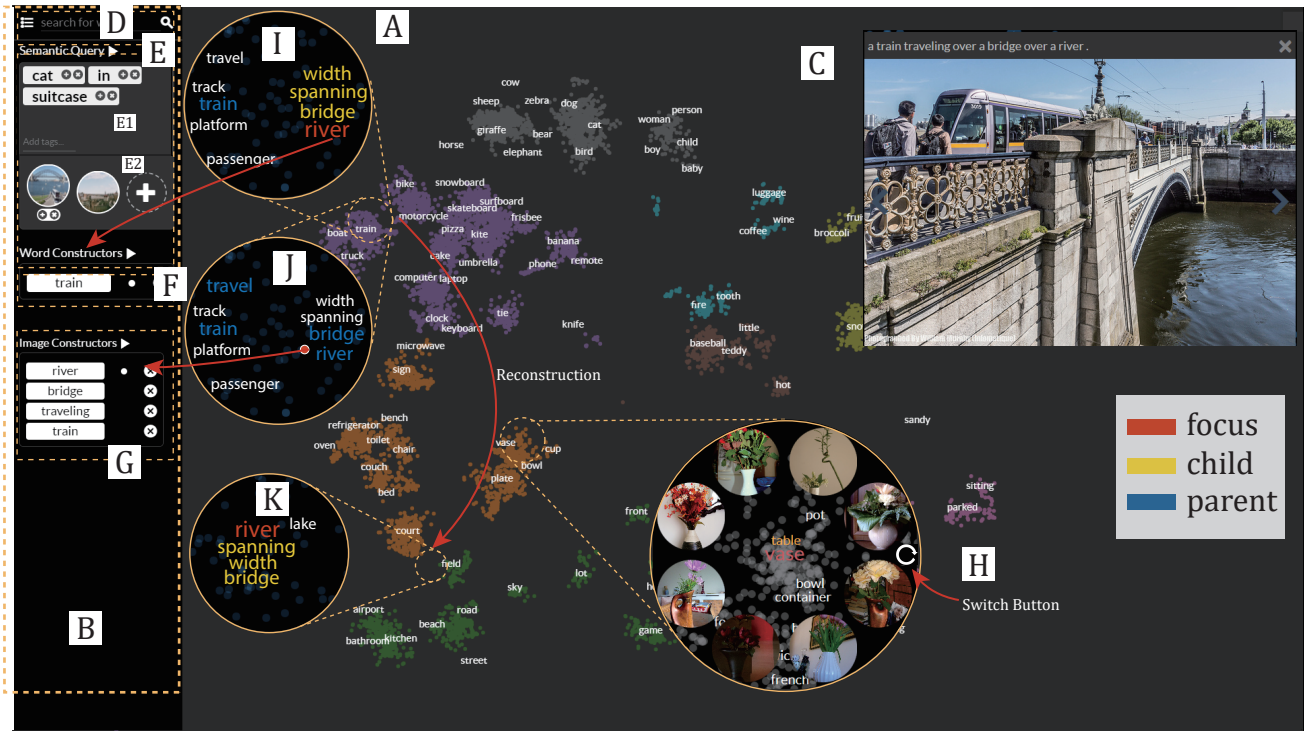


Fig. 3. The interface of the system. The system comprises of a galaxy view (A), a side panel (B), and an image browser (C). Users can locate to a word in the galaxy view by specifying in (D). Semantic query is provided in (E), where the keyword list is in (E1) and the key image list is in (E2). The word constructors are shown in F and the image constructors are shown in (G). (H) is a local area of the galaxy view. (I), (J), and (K) shows different statuses of the same local area under different interactions.

4.2.2 Reconstruct Images in Semantic Space

In this step, we attempt to reconstruct the preliminary co-embedding based on the relation between words. The relation is defined by words' co-occurrence relationships. We use the co-occurrence because it is intuitive for users and has been extensively applied [46] to reveal semantic relations. As shown in Fig. 2(E), we determine the semantic relation in a forest structure according to R1. We didn't use existing ontologies because their structures remain static and unchangeable. Moreover, multiple ontologies may contradict one another on certain aspects, as they are created in different contexts. Then, we reconstruct the position of words, where the child word's position is determined by its parent words. The word that does not have parent node just remain its original position as shown in Fig. 2(A).

Reconstructing images by words. We have detected the local semantic structures in images. Thus, we can utilize the relation between words to reconstruct the images according to their parent word. We denote the 2D positions of W_i in $\mathcal{P}^*$ as $P^*_{W_i}$. On the basis of the relation in W_{space} and the local semantic structures in the forest form, we can reconstruct the positions of the images as:

$$P^*_E = P_E - P_{Cons(E)} + P^*_{Cons(E)} \tag{11}$$

where E is an element, which can be an image or a word, and $Cons(E)$ is the constructor of the element E. The constructor is the parent node of E in our forest. Typically, the parent node of an image is the word most related to it. According to the previous step, some images may not have any parent node. We move these images into the unreconstructed list. Since each word is a root in the forest, we initially set the constructor of each word as itself. Simply reconstructing the co-embedding based on the current forest would result in a image cluster separated by different keywords as constructors. As illustrated in Fig. 2(E), there is a image cluster whose captions contain the word "dog". However, some of these images would choose "walk" as their parent, and thus the image cluster associated with "dog" is departed by "walk". Therefore, we need to introduce a process that reconstruct the word using other words.

Reconstructing words by words. In this mechanism, a word can also be reconstructed by the other words to control the reconstruction of images. First, we calculate the frequency of a word by:

$$Freq(W_i) = |\mathcal{I}_{W_i}| \tag{12}$$

The co-occurrence frequency of two words is expressed as:

$$Freq(W_i, W_j) = Freq(W_j, W_i) = |\mathcal{I}_{W_i} \cap \mathcal{I}_{W_j}| \tag{13}$$

We define the *confidence* between each pair of words as matrix $\mathbf{CF}$ where $\mathbf{CF}_{ij}$ is the confidence of word W_i can be generated by another word W_j:

$$\mathbf{CF}_{ij} = \frac{Freq(W_i, W_j)}{Freq(W_i)} \tag{14}$$

We allow the user to set a threshold, which is denoted as $MinConf$, to control the minimal confidence that a word can be generated by another word. According to the confidence, we can obtain a constructor list $\mathcal{W}_{W_i}$ (a list of potential parent nodes) of W_i. If W_j is in $\mathcal{W}_{W_i}$, then:

$$\mathbf{CF}_{ij} > max(\mathbf{CF}_{ji},\ MinConf) \tag{15}$$

For W_i, to determine its parent node, we sort the words in $\mathcal{W}_{W_i}$ according to their confidences ($\mathbf{CF}_{ix}$) and to their 2D distance from W_i. Then, we select the word that ranks first in $\mathcal{W}_{W_j}$ after sorting as the parent node of W_j. To illustrate, we construct a pair of values ($\mathbf{CF}_{ji}$, $||W_j - W_i||$) for each

word W_i in $\mathcal{W}_{W_j}$. We ensure the word W_i ranks ahead W_k if $\mathbf{CF}_{ji} < \mathbf{CF}_{jk}$ holds, or $||W_j - W_i|| < ||W_j - W_k||$ is true when $\mathbf{CF}_{ji} = \mathbf{CF}_{jk}$ holds. If $\mathcal{W}_{W_i}$ is empty, W_i has no parent node and it is its own constructor. Otherwise, we can treat the first member W_j in $\mathcal{W}_{W_i}$ as the parent of W_j, and the position of W_i is determined by the position of W_j.

As shown in Fig. 2(E), "dog" is determined to be the parent of "walk". The confidence of "dog" to "walk" is 0.75, whereas that of "happy" to "walk" is 0.25. In this figure, "happy" does not have a parent node, so in Fig. 2(F), "happy" is in its original position as shown in Fig. 2(A), whereas the position of "walk" is determined by "dog". Word "dog" remains its position and "walk" remains its relative position to that of "dog". The positions of images are determined by the locally semantic structure obtained in the first step. Since (13), (14), and (15) guarantee that $Freq(W_i) < Freq(Cons(W_i))$, there will not be any cyclic constructors. Then we can reconstruct those words whose constructors exist iteratively using (11). After reconstructing these words, we can reconstruct the images according to (11) as well. Finally, we provide a co-embedding that preserves the semantic relation globally and the visual relation locally.

5 Visual Design

In this section, we introduce the visual design and interactions of our system (Fig.3). As discussed in Section 4.2, the images are organized globally by the words, showing the semantic similarity, and distributed locally with regard to their visual similarity. However, an intuitive visual design and useful interactions are required to support efficient analysis of the images. Hence, we propose a novel visualization to provide an interactive image analysis process.

As illustrated in Fig. 3, our visualization contains several useful components. The galaxy view (Fig. 3(A)) adopts a galaxy metaphor to allow the multi-scale analysis of images. Users can zoom into the detail level of a semantic group of images to examine local patterns. Flexible query of images is supported in the side panel (Fig. 3(B)), where users can select images and words to be included and excluded(Fig. 3(E)). When users locate to an interesting word (Fig. 3(D)), its related images are displayed in a focus-plus-context manner (Fig. 3(H)). A browser (Fig. 3(C)) is provided for users to inspect the original image along with its caption.

5.1 Galaxy view

The galaxy view employs a galaxy metaphor to show the semantic summarization and semantic structure of an image collection (**R1, R2**). Organizing images according to their similarities is important for understanding the image collection [48]. Considering the projection-based method as an intuitive and concise choice for showing the similarities visually, we propose using scatterplots for the visualization. Scatter plots are regarded as basic-level visualization tools and efficient at presenting two quantitative value attributes for viewing distributions within the data. Hence, we use scatterplots to visualize both the images and words.

Galaxy metaphor. In section 3.2, we discussed the necessity of using an intuitive metaphor for visualization. We decide to employ the galaxy metaphor (**R3**) for several reasons. First, a galaxy is assembled by numerous stars and massive materials. The vastness of galaxies is appropriate for representing massive images. Second, the galaxy metaphor entails an inherent hierarchical structure (galaxy, star, planet), which can provide users with an intuitive interface to navigate and analyze the image collection at different levels of details (**R1**). Finally, the relation between stars and planets is also appropriate for representing the wordimage relation, given that different images share the same concept (word). Furthermore, using this metaphor prioritizes words (stars) over images (planets) in the visualization, and this approach encourages users to first explore the image collection through word inspection.

As shown in Fig. 3(A), we refer to each cluster as a galaxy and use color to encode these galaxies. Specifically, we refer to words as stars and images as planets in galaxies. To mitigate the overlap issue of visualizing original images, this design emphasizes words to utilize their abstractness to assist users in rapidly inspecting a large image data set. To derive the galaxies, we apply a density-based clustering method to cluster the images and words based on their 2D embeddings. We arrange images and words globally according to their semantic similarities. Thus, each galaxy can represent a set of semantically similar images (**R2**).

By identifying stars (words) that fall in different galaxies, users can quickly notice the semantic content of the image collection. Furthermore, users can detect semantically near in addition to semantically apart galaxies from estimating the distance on scatterplots. This process can help users understand the semantic structure in images, which is significant for further explorations.

5.2 Interaction

Multiple interactions are provided as follows.

Multi-scale Exploration. Due to the limited visual space, the visualization does not show all images to users at a time. To strengthen the exploration of the image collection, we employ a multi-level analysis process (**R1**). The root words of the forest structure (Section 4.2.2), considered as the most dominant concepts in the image dataset, are presented at a coarser zoom level. Moreover, users can drill down to a specific galaxy to view more detailed keywords. As discussed, we are able to group images that share similar semantic content (**R2**). For example, we place multiple images containing different species of animals together because they both share a concept, i.e., animals. However, detailed information of animals, including species distributions and their relations, are obscure at the galaxy level. To show detailed information, we allow users to zoom into the second visualization level, which is the solar level. In the solar level, each image is visualized as a planet (point). In addition, related planets (images) are shown in a focus-plus-context manner by hovering on a star (word) (Fig. 3(H)). Planets are placed radially around the star, which constitutes a typical solar system. Users can hover on a planet to show its caption, thereby verifying the correctness of the layout. The original image with high resolution (Fig. 3(C)) is shown by clicking the planet. For simplicity, we show only several related images (up to 10 images) by default. Users can click on a switch button (Fig. 3(H)) to see the rest of the images.

Layout Refinement. To help users efficiently analyze the images, we design a set of interactions to involve their domain knowledge in refining our projection layout (**R1**). When users click on an image point (Fig. 3(J)), a list of related words are shown as candidates (Fig. 3(G)). Only one of these words is the constructor of the image, which is

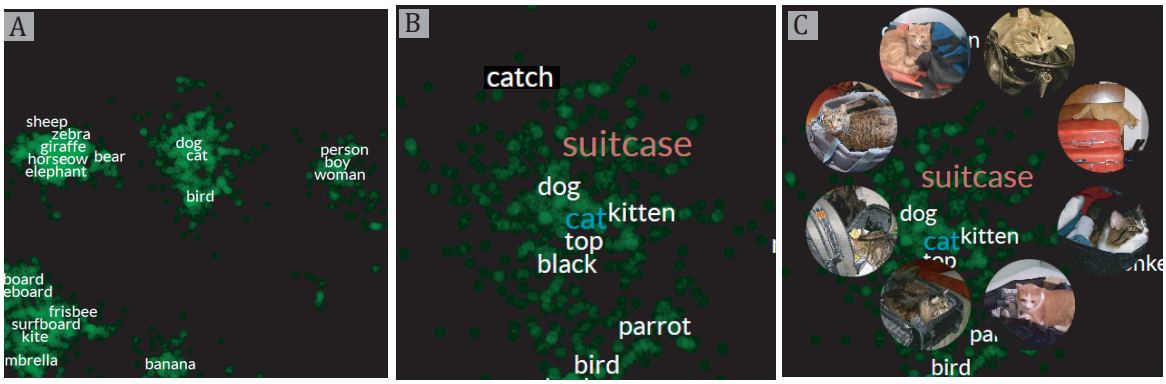


Fig. 4. The usage scenario. (A) The overview of the image collection. (B) The solar level of the galaxy about pets. (C) The relevant images about "suitcase".

highlighted by a black spot. As mentioned in Section 4.2, the position of an image is highly determined by its constructor. We typically select the most similar and the nearest one as the constructor. Depending on the task domains, users can click on other related words to choose a word that is more accurately related to the image to refine the layout.

Users can also modify the constructor of words. A constructor is created for each word to reorganize the layout of the words. As discussed in Section 4.2, this process can help prevent the separation of similar images during co-embedding. Specifically, the constructor can be interpreted as the parent word. Thus, we organize the words in a tree structure based on the semantic content of images (**R2**). However, the constructor is determined by the image captions, which are not always reliable for all domains. Therefore, we allow users to redefine the constructors of words. A constructor list of words (Fig. 3(F)) is displayed by clicking the corresponding star in the galaxy view (Fig. 3(I)). Users can manipulate the construction relations between words by reconfiguring the constructor. For example, when users remove word "train" from the constructor of word "river", the related images of word "river" are removed from the tree of word "train". Then these related images and word "river" are reorganized only according to semantic similarity between word "river" and other words. Hence, as we can see in Fig. 3, these images and word "river" are positioned to the adjacent area of word "lake" (Fig. 3(K)).

Semantic Queries. In the side panel (Fig. 3(E)), users can add related keywords and images for image query. To select the related images, users can click images in the galaxy view or the browser. Users can switch the query status for each item, either *plus* or *minus* in Fig. 3(E1, E2). The query result of images is shown in the browser. Supporting this kind of flexible query mechanism could significantly help users in understanding images.

In our system, the flexible query of images (**R4**) can be represented as an expression containing both images and words. For example, the expression *"dog"+"cat"+"image of the beach"* can represent a query of images that contain a dog and a cat in a scene similar to the image of the beach. Further, specifying an expression *"image of flower"-"red"* can represent finding images of flowers except for red flowers. We achieve flexible queries by utilizing the linear property of word embeddings. In the queries, each image is considered a document. We clean up the stop words and sum up the word embeddings to acquire the document embedding. Adding embeddings to or subtracting embeddings from the expression can obtain the vector representation of the query and search images that have captions similar to this query.

6 Usage Scenario

We present a scenario based on exploring a personal album. Bob, a fan of digital devices, is keen on capturing his life with his camera. We describe how he uses our system to manage his photos. The operations involved in this task are common for understanding general images.

Bob has taken photos for years. Consequently, he cannot clearly review what he has recorded. To have an intuitive summary, he goes to the galaxy view and soon recognizes that the overview of his photo collection is composed of several galaxies. In his observation of the stars inside the galaxies, he notices that these galaxies represent different types of photos and subjectively abstracts them as recordings of sports, person, wild animals, and pets(Fig. 4(A)). Up to now, Bob has understood the major semantic content embedded in the photos.

Bob is interested in photos of pets; thus, he drills down to the corresponding galaxy. At the solar level, as illustrated in Fig. 4(B), the photos include several kinds of pets, such as dogs, cats, and birds. Zooming into the layout, Bob finds additional descriptive words in addition to nouns, such as "swimming" (placed near "duck") and "catch" (placed near "dog"). On the basis of the positions of these stars, he speculates that they can represent the different activities of the pets. By hovering on "catch", Bob discovers a set of photos depicting specific activities, such as a dog catching a frisbee, and his speculation is therefore supported. Although the positions of most of the stars can be comprehended, a specific star called "suitcase" catches his attention because, according to common sense, the word is not semantically similar to pets. By clicking on star "suitcase", Bob finds that star "suitcase" is specified as a child of star "cat". Thus, he realizes that "cat" and "suitcase" are tightly correlated in his photos. However, Bob is confused by this pattern. He decides to further investigate the correlation. He continues inspecting the photos related to "suitcase" to explore the reason for its position. After recognizing the photos (Fig. 4(C)), he verifies the association: the photos share a common scene of a cat lying in a suitcase. Bob soon understands that cats commonly lay in suitcases. After discovering this correlation, Bob tries to collect these photos. Therefore, he utilizes the function of flexible queries to acquire fine-grained photos. In particular, he conducts in-depth queries in the left panel,

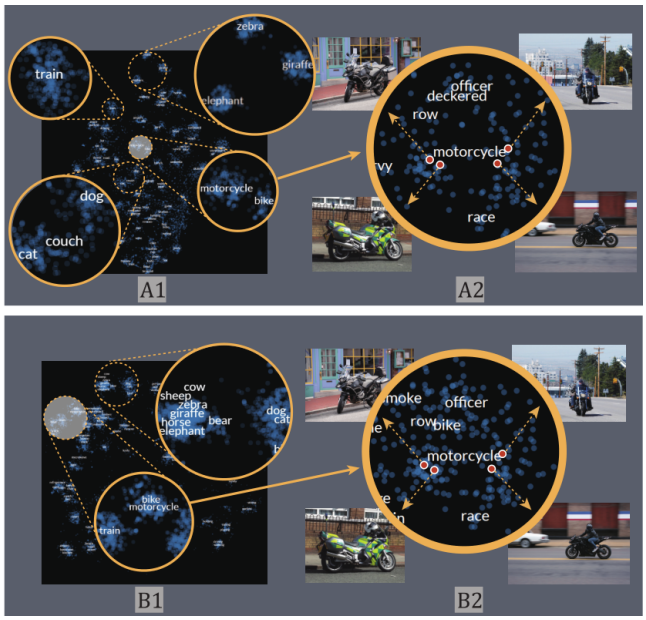


Fig. 5. The comparison of our co-embedding with previous methods. (A1) A t-SNE projection based on image embeddings. (A2) A detail layout of A1. (B1) A co-embedding projection based on image embeddings and word embeddings. (B2) A detail layout of B1.

such as "cat"+"in"+"suitcase" and "cat"+"top"+"suitcase", and successfully collects photos of cats in different motions.

Although this pattern is reasonable for Bob, he prefers to reorganize the photos of "suitcase" to the adjacent place of semantically similar words. By clicking on the removing icon of word "cat" he removes "cat" from constructor of "suitcase" and specifies word "suitcase" itself as a root word. This time, "suitcase" and related photos will be laid out only according to the semantic similarity. Therefore, Bob successfully reorganizes the photos of "suitcase" and places them in a new group containing "blender" and "purse".

7 Model Evaluation

In this section, we first present a layout comparsion to introduce the characteristics of the semantic based image layout. Later, we discuss the effect of different parameters of the co-embedding model on the image layout. Finally, we clarify the computing times of the co-embedding model.

7.1 Layout Comparison

We use an example to demonstrate the effectiveness of our co-embedding method compared with previous methods. For the comparison, we create two image layouts generated by the t-SNE (Fig. 5(A1)) and our co-embedding method (Fig. 5(B1)), respectively. The words are positioned based on the method discussed in section 4.2.1.

We first compare the projection overview. In Fig. 5(A1), the visually similar images are grouped together. As we can see, images of different animals, such as elephants, zebras, and giraffes are grouped into different clusters. However, the overview of the images still has some problems. In Fig. 5(A1), dogs and cats are positioned far away from other animals, thus breaking the semantic relation of images about animals. We speculate that because dogs and cats are pets that often appear in indoor scenes, their images are not so similar with those containing other animals, as the backgrounds are significantly different. Another example has to do with the modes of transportation. Due to the different visual appearances, various modes of transportation, such as a car, a motorcycle, and a boat, are located separately in the space. Hence, utilizing only the visual features cannot reveal the semantic relations embedded in images. As shown in Fig. 5(B1), our co-embedding method produces a more reasonable layout of images. Images are also grouped into different clusters, helping users identify their proper distribution. Further, the semantically similar images, such as the animals, modes of transportation, and indoor elements, are successfully placed together. Obviously, our co-embedding method can attain a better semantic layout of images.

Thereafter, we select an area about "motorcycle" in the overviews to compare the detail layout. In Fig. 5(A2), the images are distributed strictly according to their visual appearances. This property can assist users in identifying similar images. This property is also preserved in our projection (Fig. 5(B2)). Therefore, this verifies the capability of our co-embedding method of integrating the advantages of both the visual features and the semantic information.

7.2 Parameter Anlysis

Several parameters are involved in the co-embedding process. The parameter MiniSimi controls the sensitivity of collecting semantically similar images of a word. For the high values of MiniSimi, a few related images are detected for each word. Thus, the position of each word in the image space is determined by the limited related images that produce a dense layout of images (see Fig 6(A1)). A low MiniSimi increases the number of related images and ensures the completeness. However, according to the embedding method, it is difficult to find a position that is close to every related image. Therefore, such a difficulty would cause a loose layout of images (see Fig 6(A2)).

MaxDist controls the affected area of each word. We use red points to represent the related images of the word "plane" (see Fig 6(B1, B2)). The affected area of the word "plane" is expanded by increasing the value of MaxDist. However, the affected areas of different words would have a significant overlap with high values of the parameter MaxDist, thereby causing blurred margins among the images of different words.

MinConf controls the difficulty of detecting the co-occurrence relationships of concepts in the image data set. The high values of MinConf enable the detection of the strong correlations only, whereas the low values of MinConf would also preserve weak co-occurrence relationships. For the image layouts, this condition is reflected as increasing the values of MinConf that would reduce the number of image clusters (see Fig 6(C1, C2)) by merging several clusters.

In our system, we set MiniSimi to 0.8 to increase the accuracy of related images for each word. MaxDist is set to 0.2 to clarify the affected area of different words, thereby increasing interpretability. We believe that, by default, users are initially interested only in strong relationships between concepts, and thus, we set MinConf to 0.8.

7.3 Time Performance Analysis

We focus on analyzing the computing times of our co-embedding method because the image captioning model can be run in advance. We specify the number of images as n and the number of words as m.

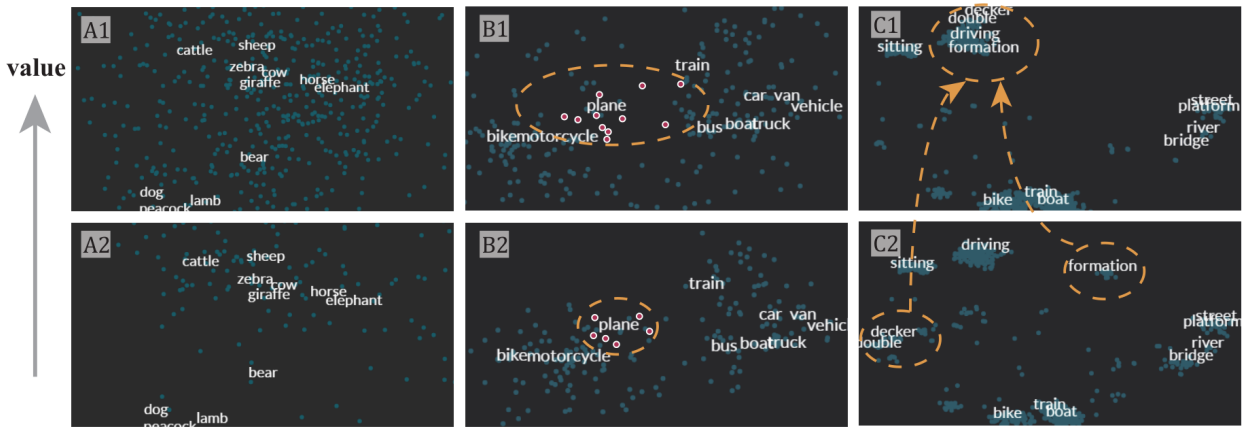


Fig. 6. The image layouts under different values of parameters. (A1) $MiniSimi = 0.8$. (A2) $MiniSimi = 0.3$. (B1) $MaxDist = 0.9$. (B2) $MaxDist = 0.1$. (C1) $MinConf = 0.8$. (C2) $MinConf = 0.5$.

In the first step, the time complexity of binding words and images is $O(mn)$ because each caption contains limited words. The time complexity of embedding words is difficult to estimate. Finding geometric median can be nearly linear, but the number of the related images of each word highly depends on the data set and parameters. Assigning a large value to MiniSimi could decrease the size of the related images of each word and accelerate the embedding process. A large value of MaxDist can also reduce computing time by decreasing the number of iterations. Hence, the worst situation of embedding words would be near $O(mnkt)$ by specifying the maximum number of related images as k and the maximum iteration number as t. Extracting the local semantic structures comprises the sorting related words for each image. Thus, the complexity is below $O(mlog(m)n)$.

In the second step, the time complexity of reconstructing images by words is $O(n)$. The time complexity of reconstructing words by words is also $O(n)$ because the co-occurrence of any two words can be computed by iterating over the captions of images. Overall, the time-consuming component of our co-embedding method is the first step.

Most of the time in our implementation was spent on the pre-processing for the co-embedding of 10000 images and approximately 600 keywords. MiniSimi was set to 0.8 and the complete co-embedding procedure took less than 3 minutes. However, a small value of MiniSimi (0.2) would increase the time cost to approximately 35 minutes.

8 Experiments

In this section, we report on two experiments conducted to test the effectiveness of our system. The first quantitatively evaluates the semantic-based image layout of our method. The second experiment evaluates the different aspects of our system, including the usefulness of the various aforementioned interactions on image analysis.

We used the MS COCO dataset [49] for the experiments. For the quantitative studies, we needed a ground truth of image information to evaluate user performance. Labeling a sizeable amount of images is a labor-intensive task; therefore, we decided to use machine learning image datasets with provided ground truths. However, many machine learning datasets collect only simple images (i.e., close-up of certain objects), which do not match real image analyses. Thus, we chose MS COCO from among different machine learning datasets because it was created by collecting complex images that depict everyday scenes.

We conducted our experiments on a desktop computer (with a 27″ screen and 1920×1080px resolution).

8.1 Experiment 1: Study of Image Layouts

To our knowledge, we are the first to develop a semantic image layout. The visual-based layout was used as the baseline, then this visual-based layout was compared with the semantic-based layout. We used t-SNE, a state-of-the-art image projection technique, to acquire the baseline. Words were positioned according to the method discussed in Section 4.2.1. Thus, the baseline layout was consistent with previous methods [8]. For testing, we designed a system prototype by removing the interactions of layout refinement and semantic queries. Users were allowed to use only zooming and basic interactions with the inspected images. A total of 12 volunteers (7 males and 5 females) participated in this study. Participants were either graduate or undergraduate students who are CS majors, and all reported normal or correct-to-normal vision.

In Section 3.1, we have described four common tasks in image analysis . Among these tasks, the performance of T1 and T3 is related to image layout, whereas those of T2 and T4 are determined by query algorithms and interactive refinements, respectively. The performance of T3 is difficult to measure, and thus, we designed the task in this experiment mainly based on T1. This task entails exploring a dataset and identifying salient object categories. We provided each participant with a set of predefined object categories and asked them to select salient categories from the set. Each participant completed this task with two image layouts separately. The accuracy of the task regarding one image layout is measured as the number of correct selections.

In this experiment, we adopted a within-subject design to compare two image layouts. The dataset cannot be repeated for each participant, and thus, we need two separate datasets. To balance the effects of the dataset and the sequence, we equally sampled all the conditions of the combination of dataset and layout, along with the sequence. We derived four conditions for each participant by denoting two image datasets as D_I and D_O and two image layouts as L_V and L_S. These conditions are as follows: $[D_I L_V, D_O L_S]$, $[D_I L_S, D_O L_V]$, $[D_O L_V, D_I L_S]$ and $[D_O L_S, D_I L_V]$. We

TABLE 1
Ground Truth

Categories of Each Partition	
Indoor Scene	Outdoor Scene
Food, Animal, Indoor Obj, Person and Accessory	Sports, Outdoor Obj, Animal, Vehicle, Person and Accessory

then divided the participants into four groups and assigned a condition to each group.

8.1.1 Procedure

We first introduced the characteristics of the two image layouts and the definition of all image categories to the participants. They were then asked to complete the task on a training dataset, which is a small set of test images that contained two image categories and approximately 800 images. The participants were trained with both the semantic-based and visual-based image layouts. During the training process, we instructed the participants to explore images and indicate their misunderstandings. Lastly, for image exploration under the condition of our specifying task, we conducted a short interview with each participant regarding the image layout they preferred. The identification result and costing time were collected for further analysis. The duration of the experiment for each participant was approximately 30 minutes.

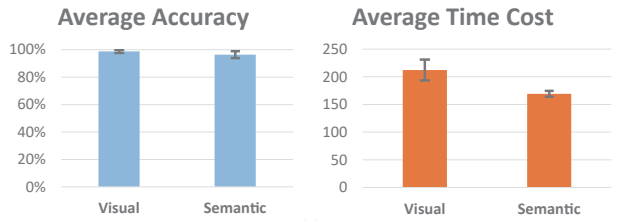


Fig. 7. Experiemnt 1: Average accuracy and time cost and their 95% confidence intervals.

8.1.2 Result

The average accuracy and costing time are presented in Fig. 7. We use Student's t test to examine the result. The left panel in Fig. 7 indicates that no significant difference ($t(11) = 0.432$, $p = 0.337$) exists between accuracies. In the right panel, however, a significant difference ($t(11) = 2.335$, $p = 0.020$) is observed between the average costing time for the two image layouts. The result is consistent with our observations. We did not establish a time limit for the task, and thus, most participants chose to thoroughly explore images until they were sure of their answers. Therefore, they achieved highly accurate results. Nevertheless, the finishing time for the two image layouts differed. During the experiment, we observed that participants were likely to exhibit more interactions with the inspected images when using the visual-based image layout. In particular, five participants noted that realizing and utilizing semantic similarities are easier for them. We speculate that the semantic similarity can be more efficiently used as it is more close to human while the visual similarity is more close to machine.

We found additional reasons for the outcomes from the interviews. The characteristic of clustering semantically similar images and labels could help them reduce summarization time given that they were familiar with the concept of semantics and were certain of the content of each cluster without considerable interactions of inspecting images. By contrast, the concept of visually similar images will be fuzzy to the participants. Whether two images should be visually similar and the degree of visual similarity remain unclear to the participants, thereby causing them to continually inspect images to confirm their content.

TABLE 2
Tasks

T1	Summarize the image collection.	R2
T2	Summarize the image cluster containing animals.	R2
T3	Find an image containing the cat.	R4
T4	Find an image containing both the cat and the dog.	R4
T5	Find word "ocean". Check the related images and find an image containing the seagull. Revise the constructor of this image.	R1
T6	Find word "ear". Check the related images and find what they are about. Revise the constructor of "ear" to make these images located to a more reasonable place.	R1

8.2 Experiment 2: Study of System Usability

In this experiment, we tested the effectiveness of our system on image analysis. We randomly sampled 10000 images from the MSCOCO as testing images. To ensure that the users would be engaged in the scenario of visual analysis of images, we designed 6 tasks (Table. 2) on the basis of our design rationales. These tasks were designed to instruct the users to fully exploit different functions of our system. We also developed a questionnaire (Table. 3) based on the tasks. The questionnaire is a seven-point Likert scale (1 - strongly disagree, 7 - strongly agree). With the tasks, we expected to collect comprehensive feedbacks and suggestions of users. A total of 15 volunteers (10 males and 5 females, with an average age of 21.2 years) participated in this study.

8.2.1 Procedure

At the beginning of the experiment, we briefly introduced our system to each participant. Then, the participants were asked to complete the tasks in Table. 2. After finishing all the tasks, the participants were given a chance to freely explore the whole image dataset. Then, they were asked to fill out the questionnaire (Table. 3) to evaluate our system. Finally, we conducted an interview with each participant to collect some feedback on our system. The typical time for the user study, including the questionnaire and the interview, was approximately 40 minutes per participant.

8.2.2 Result

The participants were able to easily complete the tasks. Given that most of these tasks are either open questions or with only one solution, we did not record the completion time and the accuracy of each task. Here we focus on the questionnaire ratings and user feedbacks.

The ratings of questionnaires are reported in Fig. 8. Overall, the users agreed (gave an average rating approximate to 6) that our system is easy to learn and to use, the layout is reasonable and helpful, and the design is intuitive and aesthetically pleasing. Among all the ratings, the variances of Q2 and Q5 are comparatively large. This

TABLE 3
Questionnaire

Q1	Our system is easy to learn
Q2	Our system is easy to use
Q3	The overview of the image collection is reasonable
Q4	The overview of the image is useful
Q5	The local structures of the images are reasonable
Q6	The local structures of the images are helpful.
Q7	The result of the semantic queries is reasonable.
Q8	The result of the semantic queries is useful.
Q9	The visual design is intuitive
Q10	The visual design is aesthetic.

can be attributed to the fact that two users gave low marks to these two questions respectively. One user gave 3 points on Q2. His remark was that the search engine made him confused, and it would be better if the system could provide more instructions. The other user gave 4 points on Q5. His comment was "it would be better to use lines to show the construction relationship."

Fig. 8. Experiment 2: Average ratings and 95% confidence intervals.

8.3 User Feedback

We summarize the user feedback below.

Semantic Layout. The users were impressed by the semantic layout of images. They thought that the layout is useful for them to explore the massive images. But one of the users mentioned that several images were positioned by mistake. Other users provided some suggestions for the co-embedding. One of the users with research experience in visualization suggested that the co-embedding process could be extended to a progressive manner. He also mentioned that it would be more helpful if users can specify words of interest in the co-embedding. Further, by progressively adding or deleting words from the co-embedding, the co-embedding result can be more transparent for users to comprehend the semantic information in images.

Visual Design. Most of the users agreed that our visual design could intuitively convey the semantic information. Given that we use the scatterplot as our visualization, they found it easy for them to recognize the images and their similarities. However, one user indicated that the estimation of image similarities was not so easy, as the visualization can be explored in different scales, which sometimes hampered his estimation of the distances. Another user mentioned that it would be more useful if the images can be visualized temporally for the task of image management. In addition, three users made comments about the color. One of them thought the color is appropriate, which provided her an impression of the starry sky. Two of them thought the color is a little confusing as it did not well distinguish between different semantic clusters. Due to the limitation of our clustering method, however, some semantically similar images may be assigned to different groups. Moreover, one user thought that it was not efficient to show all the text in the same size. He hoped that important words could be highlighted by a larger text to attract his attention.

Semantic Summarization. Although we selected the most representative words for the visualization, three of our users commented that there were still too many words in the overview, imposing them to a large scale of information that they could not quickly comprehend. One of them suggested that adding simple sketches could help her efficiently understand the semantic content. Two of them suggested conducting a summarization of similar words to produce a more high-level concept, such as animals, food, and sports. They shared that only showing these highly-summarized words in the overview could facilitate their analysis.

8.4 Discussion

Significance. With the rapid development of deep learning, an increasing amount of semantic content embedded in images can be conveyed in a readable manner. Due to the generality and abstractness of semantics, visualizing images with the aid of semantic information is a probable choice for analysis. In this work, we present our initial attempt to utilize high-level semantic information to produce a co-embedding of images and words. The evaluation demonstrates the effectiveness of combining our co-embedding model and an interactive visualization for analyzing image content. Compared with traditional methods, our method has the capability of maintaining both the semantic and visual similarities of the images, thus providing users with a new approach for exploring complex image contents. The semantic layout of images allows users to detect semantically similar images, that share an identical concept. Further, our co-embedding can characterize the semantic relation in a specific image collection based on the semantic information. The flexibility in our co-embedding also enables users to refine the image layout iteratively, thereby providing users with an interface, with which they can integrate their domain knowledge in the analysis.

Applicability. From revealing the semantic information of images, our method can be applied to facilitate the utilization of images in further analyses. The semantic information of images can be highly effective in many cases. For example, companies would like to collect and analyze user feedback on the products. However, except for the text, user feedback also contains rich images, which often show the problem product. These images are beneficial complements to the feedback analysis. Our method can be integrated into current analysis tools to conduct analyses involving the information contained in texts and images, such as linking the sentiment analysis of text with user-provided images to discover users' primary concern of the products.

Limitations. Our method has three limitations. The first limitation is the performance of the captioning model. Due to the non-optimized code and the use of a low-end PC (Tesla K20), extracting captions from 10,000 images consumes 11 hours. To resolve this issue, we can use advanced GPUs and optimize the code to accelerate model performance. We believe that the time cost can be reduced into less than 1 hour. In addition to the speed issue, the quality of the captions is highlighted. Although this model is a state-of-the-art technique for computer vision, it may not always produce reasonable descriptions. Furthermore, the captioning model is currently incapable of generating highly specific descriptions of images, thereby limiting the

richness of the semantic information. However, we believe that the fast-developing deep learning models can produce increasingly useful and concrete semantic information in the future. Therefore, our co-embedding method can be integrated with the future novel model coherently and uncover more insights from the images. The second limitation comes from our co-embedding model. As shown in Section 4.2, our co-embedding introduces several parameters. These parameters, however, highly depend on manual adjustment to attain a reasonable layout. At present, we determine the parameters by iterative experiments. In the future, we plan to design a new method which can automatically or semi-automatically determine the parameters. The design of co-embedding also posts challenges to streaming applications. This is because the t-SNE projection, which we use in preprocessing, does not support incremental updates. Substituting some incremental projections for t-SNE may resolve the problem. However, this would involve a trade-off between layout accuracy and speed. The third limitation is the removal of the sentence structures of captions. As illustrated in Fig.1, we derived captions from images and replaced them with keywords. Although we could perform a concept-level analysis by combining object labels and other descriptive words, the sentence structure has a high potential for facilitating image analysis.

9 Conclusion

This paper introduces a visual analytic system for analyzing large image collections supported with the semantic information of images. We apply an image captioning model to automatically extract descriptive captions from images. A novel co-embedding model is introduced to project images and the associated semantic keywords to the same 2D space for a semantic-based exploration. The system employs a galaxy-based design to characterize the 2D projection, thereby providing a multi-scale visualization that shows a semantic summary in addition to a detailed illustration of the images. Multiple interactions are proposed to involve the users domain knowledge in the co-embedding process to refine the projection layout. In the future, we plan to apply a more powerful image captioning model to detect more accurate and detailed semantic information embedded in images. In order to support progressively adding or deleting words, we will develop or adopt appropriate progressive t-SNE to accelerate our co-embedding model. We will also attempt to integrate natural language processing methods to provide more useful interactions for users to investigate the relationship between keywords and images.

Acknowledgment

The work was supported by National 973 Program of China (2015CB352503), NSFC (61761136020, 61502416), NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization (U1609217), Zhejiang Provincial Natural Science Foundation (LR18F020001) and the 100 Talents Program of Zhejiang University. This project is also partially funded by Microsoft Research Asia.

References

[1] W. Plant and G. Schaefer, "Visualisation and browsing of image databases," in *Multimedia Analysis, Processing and Communications*, 2011, pp. 3–57.

[2] B. B. Bederson, "PhotoMesa: a zoomable image browser using quantum treemaps and bubblemaps," in *Proceedings of the ACM symposium on User Interface Software and Technology*, 2001, pp. 71–80.

[3] Y. Gu, C. Wang, J. Ma, R. J. Nemiroff, and D. L. Kao, "iGraph: a graph-based technique for visual analytics of image and text collections," in *Proceedings of IS&T/SPIE Conference on Visualization and Data Analysis*, ser. SPIE Proceedings, vol. 9397, 2015, p. 939708.

[4] G. P. Nguyen and M. Worring, "Interactive access to large image collections using similarity-based visualization," *Journal of Visual Languages and Computing*, vol. 19, no. 2, pp. 203–224, 2008.

[5] D. Ryu, W. Chung, and H. Cho, "PHOTOLAND: a new image layout system using spatio-temporal information in digital photos," in *Proceedings of the ACM Symposium on Applied*, 2010, pp. 1884–1891.

[6] L. Tan, Y. Song, S. Liu, and L. Xie, "ImageHive: Interactive content-aware image summarization," *IEEE Computer Graphics and Applications*, vol. 32, no. 1, pp. 46–55, 2012.

[7] M. Worring, D. Koelma, and J. Zahálka, "Multimedia Pivot Tables for Multimedia Analytics on Image Collections," *IEEE Transactions on Multimedia*, vol. 18, no. 11, pp. 2217–2227, 2016.

[8] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, W. Ribarsky, and M. O. Ward, "Semantic image browser: Bridging information visualization with automated intelligent image analysis," in *Proceedings of IEEE VAST*, 2006, pp. 191–198.

[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge," *Computing Research Repository*, vol. abs/1609.06647, 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of International Conference on Neural Information Processing Systems*, 2012, pp. 1106–1114.

[11] R. Bernardi, R. Çakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures (extended abstract)," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 2017, pp. 4970–4974.

[12] P. Agarwal and A. Skupin, *Self-Organising Maps: Applications in Geographic Information Science*. Wiley, 2008.

[13] S. Krishnamachari and M. Abdel-Mottaleb, "Image Browsing using Hierarchical Clustering," in *Proceedings of IEEE Symposium on Computers and Communications*, 1999, pp. 301–307.

[14] T. Liu, J. Wang, J. Sun, N. Zheng, X. Tang, and H. Shum, "Picture Collage," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1225–1239, 2009.

[15] M. Crampes, J. de Oliveira-Kumar, S. Ranwez, and J. Villerd, "Visualizing social photos on a hasse diagram for eliciting relations and indexing new photos," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 985–992, 2009.

[16] M. Worring and D. C. Koelma, "Insight in Image Collections by Multimedia Pivot Tables," in *Proceedings of ACM International Conference on Multimedia Retrieval*, 2015, pp. 291–298.

[17] A. Thudt, U. Hinrichs, and S. Carpendale, "The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries through Information Visualization," in *CHI Conference on Human Factors in Computing Systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012*, 2012, pp. 1461–1470.

[18] K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," in *Proceedings of IEEE PacificVis*, 2015, pp. 117–121.

[19] C. Görg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. T. Stasko, "Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 10, pp. 1646–1663, 2013.

[20] M. Krstajic, E. Bertini, and D. A. Keim, "Cloudlines: Compact display of event episodes in multiple time-series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2432–2439, 2011.

[21] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. A. Keim, "Eventriver: Visually exploring text collections with temporal references," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 1, pp. 93–105, 2012.

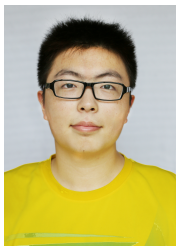
[22] F. Heimerl, Q. Han, S. Koch, and T. Ertl, "Citerivers: Visual analytics of citation patterns," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 190–199, 2016.

[23] J. Alsakran, Y. Chen, Y. Zhao, J. Yang, and D. Luo, "STREAMIT: dynamic visualization and interactive exploration of text streams," in *Proceedings of IEEE PacificVis*, 2011, pp. 131–138.

[24] K. Koh, B. Lee, B. H. Kim, and J. Seo, "Maniwordle: Providing flexible control over wordle," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1190–1197, 2010.
[25] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, and B. Guo, "TopicPanorama: A full picture of relevant topics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 12, pp. 2508–2521, 2016.
[26] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky, "HierarchicalTopics: Visually exploring large text collections using topic hierarchies," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2002–2011, 2013.
[27] M. Brehmer, S. Ingram, J. Stray, and T. Munzner, "Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2271–2280, 2014.
[28] M. Kim, K. Kang, D. G. Park, J. Choo, and N. Elmqvist, "TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 151–160, 2017.
[29] S. Gad, W. Javed, S. Ghani, N. Elmqvist, E. T. Ewing, K. N. Hampton, and N. Ramakrishnan, "Themedelta: Dynamic segmentations over temporal topic models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 5, pp. 672–685, 2015.
[30] J. A. Wise, "The ecological approach to text visualization," *Journal of the American Society for Information Science*, vol. 50, no. 13, pp. 1224–1233, 1999.
[31] J. T. Stasko, C. Görg, and Z. Liu, "Jigsaw: supporting investigative analysis through interactive visualization," *Information Visualization*, vol. 7, no. 2, pp. 118–132, 2008.
[32] I. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.
[33] C. Bentley and M. O. Ward, "Animating multidimensional scaling to visualize n-dimensional data sets," in *Proceedings of IEEE InfoVis*, 1996, pp. 72–73.
[34] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, 2002, pp. 833–840.
[35] J. Choo, C. Lee, C. K. Reddy, and H. Park, "UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1992–2001, 2013.
[36] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[37] D. R. Hardoon, S. Szedmák, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.
[38] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, 2014, pp. 7–16.
[39] J. Choo, S. Bohn, G. Nakamura, A. M. White, and H. Park, "Heterogeneous data fusion via space alignment using nonmetric multidimensional scaling," in *Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012.*, 2012, pp. 177–188.
[40] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. C. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.
[41] S. Santini and R. C. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 871–883, 1999.
[42] J. Zahálka and M. Worring, "Towards interactive, intelligent, and integrated multimedia analytics," in *Proceedings of IEEE VAST*, 2014, pp. 3–12.
[43] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li, "Large-scale video classification with convolutional neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.
[44] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Convolutional neural network committees for handwritten character classification," in *Proceedings of International Conference on Document Analysis and Recognition*, 2011, pp. 1135–1139.
[45] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015.
[46] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Computing Research Repository*, vol. abs/1301.3781, 2013.
[47] J. Dykes, A. M. MacEachren, and M.-J. Kraak, *Exploring geovisualization*. Elsevier, 2005.
[48] K. Rodden, W. Basalaj, D. Sinclair, and K. R. Wood, "Does organisation by similarity assist image browsing?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2001, pp. 190–197.
[49] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *The European Conference on Computer Vision*, 2014, pp. 740–755.

Xiao Xie received his BS in software engineering from the Sun Yat-sen University, China in 2015. He is currently a Ph.D. student in the State Key Lab of CAD&CG, Zhejiang University in China. His research interests are in multimedia visualization and visual analytics.

Xiwen Cai received his B.Sc. in Psycology from the Zhejiang University, China in 2014. He is currently a master student at the State Key Lab of CAD&CG, Zhejiang University. His research interests include human computer interaction and visualization.



Junpei Zhou graduated from Zhejiang Jinhua First Middle School in 2014. He is currently studying in the Department of Computer Science and Technology at Zhejiang University as a senior student and anticipates to receive bachelor's degree in 2018. His research interests include data visualization, data mining, and machine learning.

Nan Cao is a professor at Tongji University. He is also the founder and director of Intelligent Big Data Visualization (iDV^x) Lab. His research interests include data visualization, visual analysis, and data mining. He creates novel visual analysis techniques for supporting anomaly detection in complex (i.e., big, dynamic, multivariate, heterogeneous, and multi-relational) data.

Yingcai Wu is a ZJU100 Young Professor at the State Key Lab of CAD&CG, Zhejiang University. His main research interests are in visual analytics and information visualization, with focuses on user behavior analysis, urban informatics, social media analysis, and sports analytics. He received his Ph.D. degree in Computer Science from the Hong Kong University of Science and Technology. Prior to his current position, Dr. Wu was a postdoctoral researcher in the University of California, Davis from 2010 to 2012, and a researcher in Microsoft Research Asia from 2012 to 2015. For more information, please visit http://www.ycwu.org.